Human-in-the-Loop Imitation Learning using Remote Teleoperation

Ajay Mandlekar, Danfei Xu, Roberto Martín-Martín, Yuke Zhu, Li Fei-Fei, Silvio Savarese

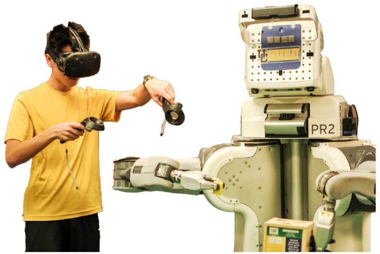
Presenter: Rutav Shah

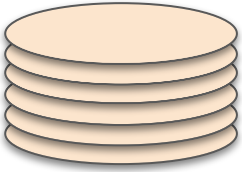
10/18/2022

Imitation learning in Humans

Imitation learning in Robots (Behavior Cloning)

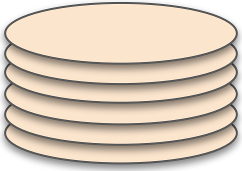
Collecting expert demonstrations



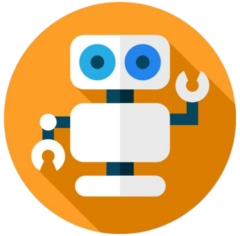


Expert Demos

Learning from demonstrations



Expert Demos



** Typical behavior cloning setup

Zhang, Tianhao, et al. "Deep imitation learning for complex manipulation tasks from virtual reality teleoperation." ICRA. IEEE, 2018.

Covariate shift: The hard regime

1. Test distribution is different from training distribution

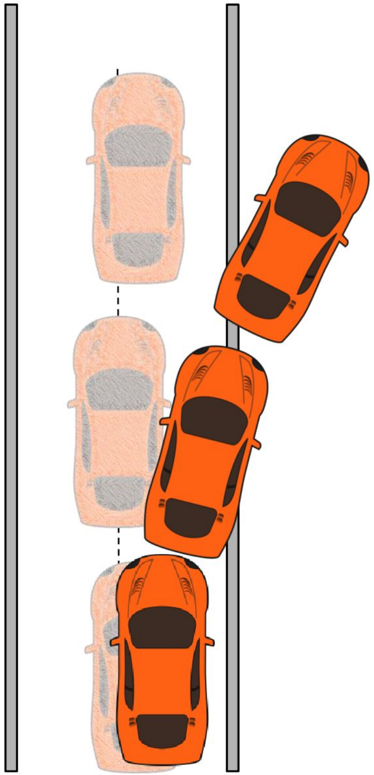
2. Compounding errors

3. **Imitation Learning**

   Train/test data are not i.i.d.

   If expected training error is ϵ
   Expected test error after T decisions
   is up to $\boxed{T^2 \epsilon}$

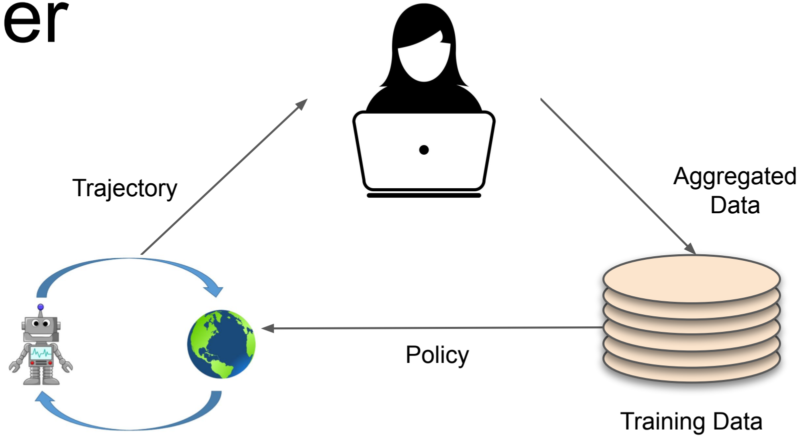
   Errors compound

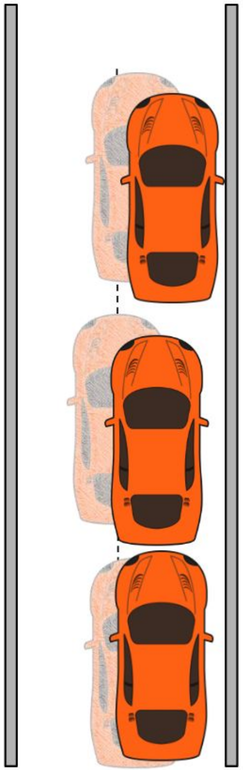


Spencer, Jonathan, et al. "Feedback in imitation learning: The three regimes of covariate shift." arXiv preprint arXiv:2102.02872 (2021).

http://www.cs.toronto.edu/~florian/courses/imitation_learning/lectures/Lecture1.pdf

DAgger





Imitation Learning via DAgger

Train/test data are not i.i.d.

If expected training error on aggr. dataset is ϵ
Expected test error after T decisions is

$$\boxed{O(T\epsilon)}$$

Errors do not compound

http://www.cs.toronto.edu/~florian/courses/imitation_learning/lectures/Lecture1.pdf

Problems in DAgger

1. Estimating correct actions

2. Relabelling entire trajectories

3. A 30-second manipulation task with 20hz robot control

    a. 30x20 = 600 state relabelling per trajectory

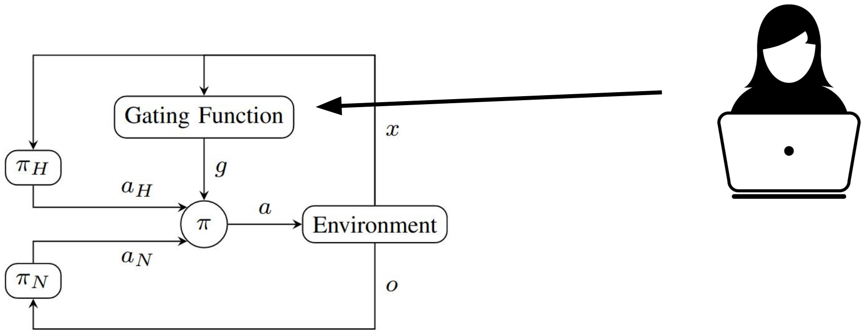
4. Unsafe

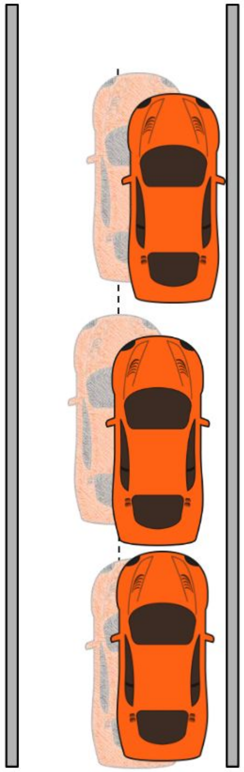
Algorithm 1 DAgger

1: $D = \{(s, a)\}$ initial expert demonstrations
2: $\theta_1 \leftarrow$ train learner's policy parameters on D
3: **for** $i = 1 \ldots N$ **do**
4: Execute learner's policy π_{θ_i}, get visited states $S_{\theta_i} = \{s_0, \ldots, s_T\}$
5: $\boxed{\text{Query the expert at those states to get actions } A = \{a_0, \ldots, a_T\}}$
6: Aggregate dataset $D = D \cup \{(s, a) \mid s \in S_{\theta_i},\ a \in A\}$
7: Train learner's policy $\pi_{\theta_{i+1}}$ on dataset D
8: Return one of the policies π_{θ_i} that performs best on validation set

Ross, Stéphane, Geoffrey Gordon, and Drew Bagnell. "A reduction of imitation learning and structured prediction to no-regret online learning." *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2011

HG-DAgger





- Intervene when necessary
- Significantly Reduces human relabelling effort

Kelly, Michael, et al. "HG-DAgger: Interactive imitation learning with human experts." *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019.

Problems in HG-DAgger

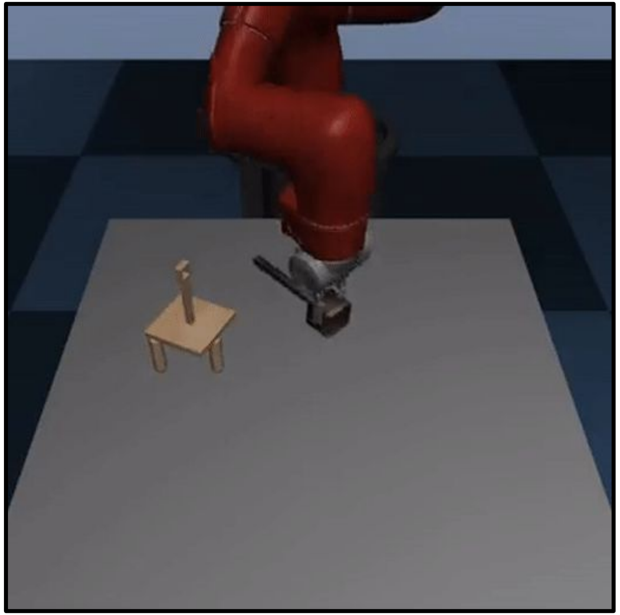
1. HG-DAgger throws away the robot sampled trajectories
2. Behavior of the agent changes significantly after training on the new dataset
3. Limited to driving scenarios

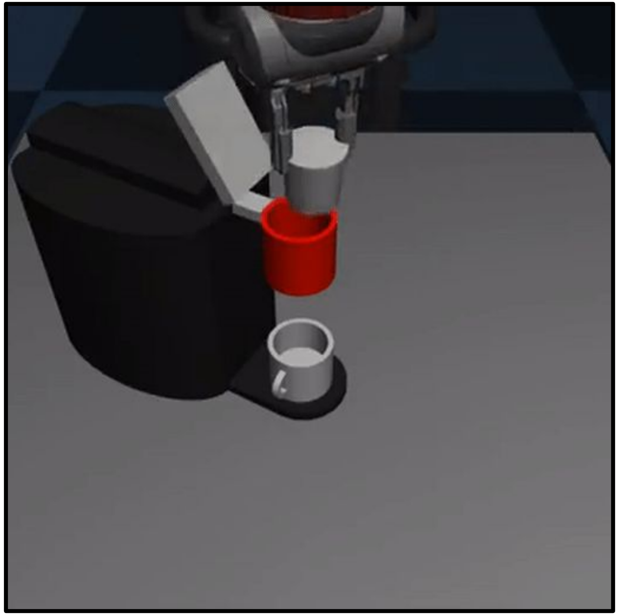
Algorithm 1 HG-DAGGER

procedure HG-DAGGER($\pi_H, \pi_{N_1}, \mathcal{D}_{BC}$)
  $\mathcal{D} \leftarrow \mathcal{D}_{BC}$
  $\mathcal{I} \leftarrow []$
  for epoch $i = 1 : K$
    for rollout $j = 1 : M$
      for timestep $t \in T$ of rollout j
        if expert has control
          record expert labels into $\mathcal{D}_j$
        if expert is taking control
          record doubt into I_j
      $\mathcal{D} \leftarrow \mathcal{D} \cup \mathcal{D}_j$
      append $\mathcal{I}_j$ to $\mathcal{I}$
    train $\pi_{N_{i+1}}$ on $\mathcal{D}$
  $\tau \leftarrow f(\mathcal{I})$
  return $\pi_{N_{K+1}}, \tau$

Kelly, Michael, et al. "HG-DAgger: Interactive imitation learning with human experts." *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019.

Robot manipulation!

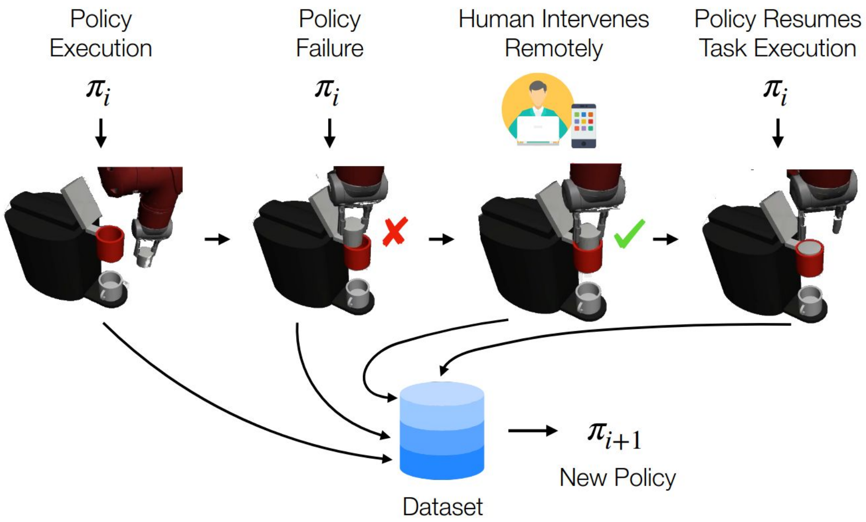




- Bottleneck regions are much more difficult to traverse

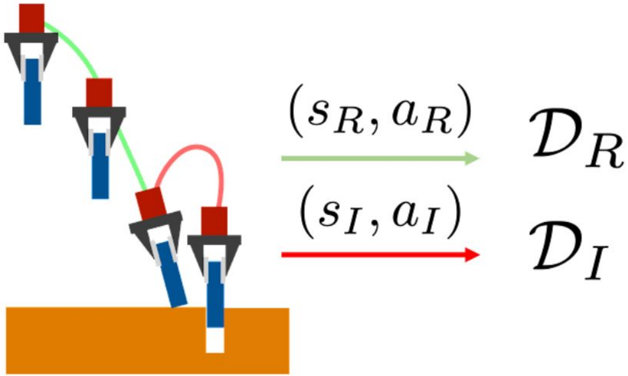
Proposed Idea

The human intervened trajectories are informative about both ***where*** task bottlenecks occur and ***how*** to traverse them.



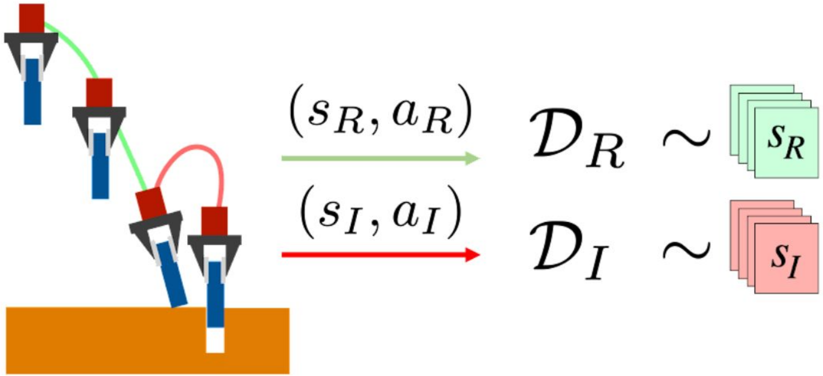
- Don’t throw away any information!

Methodology - Intervention Weighted Regression (IWR)



Red: Human intervened data
Green: Robot sampled trajectory

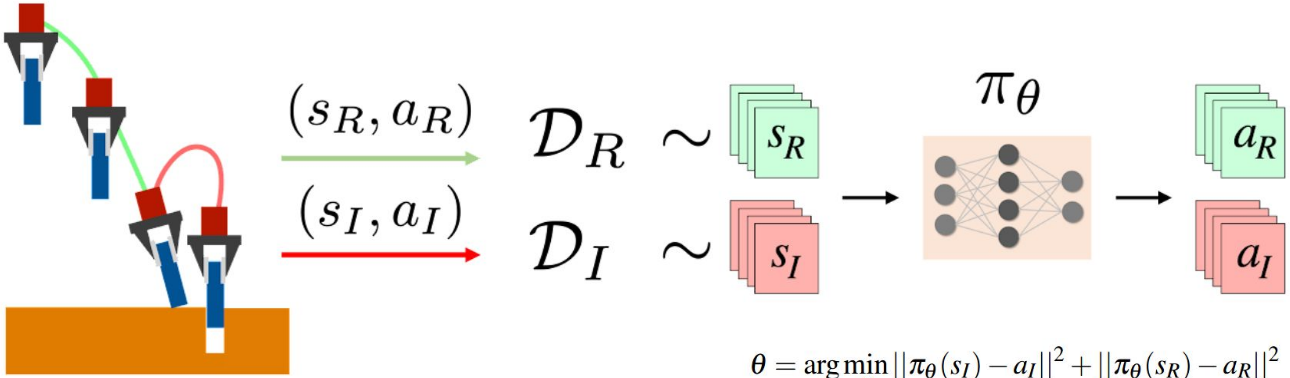
Methodology - Intervention Weighted Regression (IWR)



Sample equal batch size

Red: Human intervened data
Green: Robot sampled trajectory

Methodology - Intervention Weighted Regression (IWR)



$$\theta = \arg\min_\theta ||\pi_\theta(s_I) - a_I||^2 + ||\pi_\theta(s_R) - a_R||^2$$

Why equal batch size?

- equal size batches re-weights the data distribution (??)
  - Intervention actions demonstrate bottleneck traversal
  - Robot sampled data keeps the policy close to previous policy

Methodology - Mathematical Grounding

$$J(\theta) = \mathbb{E}_{\boxed{\pi_\theta}}[R(\tau)]$$

Variational Lower Bound

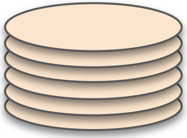
$$J(\theta) = \mathbb{E}_{\boxed{q(\tau)}}[R(\tau) + \log\ p_{\pi_\theta}(\tau) - \log\ q(\tau)]$$

Human intervened distribution
+ Robot on policy samples

Optimization - EM Algorithm

$$q(\tau) = \arg\max_q \mathbb{E}_{q(\tau)}[\log R(\tau)] - KL[q(\tau)\,||\,p_{\pi_\theta}(\tau)]$$

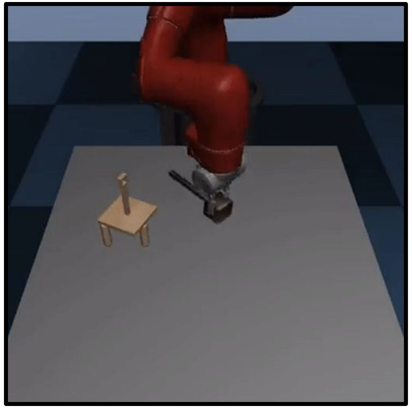
$$\theta = \arg\max_\theta \mathbb{E}_{(s,a)\sim q(\tau)}[\log \pi_\theta(a|s)]$$

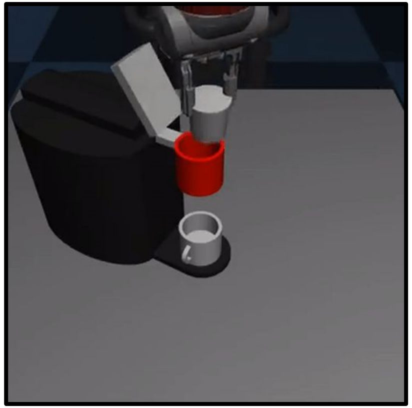


Aggregated Data

A. Abdolmaleki, J. T. Springenberg, Y. Tassa, R. Munos, N. Heess,and M. Riedmiller, “Maximum a posteriori policy optimisation”,arXiv preprint arXiv:1806.06920, 2018.

Experimental Setup





Tested on simulated environment

Remote RoboTurk system

Results

TABLE I: Single-Operator Results on the Threading Task

Model	Round 1	Round 2	Final
Base	-	-	58.0 ± 9.2
Full Demos	-	-	76.7 ± 2.3
HG-DAGGER	57.3 ± 9.5	62.7 ± 5.0	75.3 ± 8.1
IWR-NB	76.0 ± 6.9	72.0 ± 3.5	74.7 ± 1.2
IWR (Ours)	**84.0 ± 5.3**	**90.7 ± 3.1**	**87.3 ± 5.0**

IWR (Ours)	Samples equal batch size from human and robot data
IWR-NB	Mixes both the robot and human data, then sample
HG-Dagger	Throws away robot samples from human intervened traj
Full Demos	No human intervention

Results

- Train policy from scratch using the data collected by each method

TABLE III: Single-Operator Comparison across Final Threading Datasets Collected by Each Method

	Final Dataset		
Model	HG-DAGGER	IWR-NB	IWR (Ours)
HG-DAGGER	75.3 ± 8.1	72.0 ± 5.3	81.3 ± 4.2
IWR-NB	80.0 ± 1.4	74.7 ± 1.2	86.0 ± 4.0
IWR (Ours)	$\mathbf{87.3 \pm 6.4}$	$\mathbf{84.7 \pm 6.4}$	$\mathbf{87.3 \pm 5.0}$

TABLE IV: Multi-Operator Comparison across Final Coffee Machine Datasets Collected by Each Method

	Final Dataset	
Model	HG-DAGGER	IWR (Ours)
HG-DAGGER	69.6 ± 10.1	71.6 ± 16.1
IWR (Ours)	$\mathbf{85.6 \pm 6.5}$	$\mathbf{87.5 \pm 9.4}$

IWR (Ours)	Samples equal batch size from human and robot data
IWR-NB	Mixes both the robot and human data, then sample
HG-Dagger	Throws away robot samples from human intervened traj
Full Demos	No human intervention

Results

- Average results across three different operators

TABLE II: Multi-Operator Results on the Coffee Machine Task

Model	Round 1	Round 2	Final
Base	-	-	52.0 ± 3.5
Full Demos	-	-	64.9 ± 8.3
HG-DAGGER	70.2 ± 15.3	71.1 ± 9.7	69.6 ± 10.1
IWR (Ours)	$\mathbf{79.6 \pm 8.9}$	$\mathbf{79.5 \pm 11.7}$	$\mathbf{87.5 \pm 9.4}$

IWR (Ours)	Samples equal batch size from human and robot data
IWR-NB	Mixes both the robot and human data, then sample
HG-Dagger	Throws away robot samples from human intervened traj
Full Demos	No human intervention

Critiques

1. Results only on simulator not on real world tasks! **Covid :(**

2. Not convinced that ***where*** and ***how*** the bottleneck occurs has been fully addressed

3. No information about the percentage of times human had to intervene per trajectory per round

4. What if human makes an error while executing the task? Robustness to such errors?

5. **Full demos** is not a convincing baseline
   a. **Full demos** has (**30 x traj_len**) human samples
   b. **IWR(Ours)** has (**30 x traj_len + no. of intervention**) human samples

6. The comparisons are a bit inconsistent for e.g **IWR(NB)** is not compared with in Table 2 and Table 4. Best guess: Too much human annotation per seed per algorithm.

7. No access to codebase or data - https://sites.google.com/stanford.edu/iwr

Summary

- Learning from human demonstrations
  - Effective
  - Human centric world
- Human in the loop: Tackles covariate shift while minimizing human effort
- Key Takeaway: The human intervened trajectories are informative about both ***where*** task bottlenecks occur and ***how*** to traverse them.
- Demonstrates strong results on simulated environments

Extended Readings

1. Classical: Navlab 1 (1986-1989); Navlab 2 + ALVINN (1989-1993)
2. DAgger: A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning
3. Feedback in Imitation Learning: The Three Regimes of Covariate Shift - Categorizes the compounding error problem in three categories and possible solution in each one of them.
4. DART: Noise Injection for Robust Imitation Learning - Injects noise while training instead of intervention.
5. Comparing Human-Centric and Robot-Centric Sampling for Robot Deep Learning from Demonstrations - Compares human demonstrations and data collected in DAgger style.
6. Learning from Interventions Human-robot interaction as both explicit and implicit feedback - Similar idea of using interventions but instead learning constraints on the value function.
7. Robot Learning on the Job: Human-in-the-Loop Manipulation and Learning During Deployment **(ICRA 2023) -** Makes better use of the interventions made by humans
8. Imitation learning: A series of Deep Dives – Short Youtube series by Sanjiban Choudhary

Questions?

Open Questions

1. Is human-in-the-loop really scalable?

2. Can safety be learned from more sparse (or no) feedback from human?

3. Threat to optimality by using behavior cloning and/or human-in-the-loop?